

10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
160
170
180
190
200
210
220
230
240
250
260
270
280
290
300
310
320
330
340
350
360
370
380
390
400
410
420
430
440
450
460
470
480
490
500
510
520
530
540
550
560
570
580
590
600
610
620
630
640
650
660
670
680
690
700
710
720
730
740
750
760
770
780
790
800
810
820
830
840
850
860
870
880
890
900
910
920
930
940
950
960
970
980
990
1000
1010
1020
1030
1040
1050
1060
1070
1080
1090
1100
1110
1120
1130
1140
1150
1160
1170
1180
1190
1200
1210
1220
1230
1240
1250
1260
1270
1280
1290
1300
1310
1320
1330
1340
1350
1360
1370
1380
1390
1400
1410
1420
1430
1440
1450
1460
1470
1480
1490
1500
1510
1520
1530
1540
1550
1560
1570
1580
1590
1600
1610
1620
1630
1640
1650
1660
1670
1680
1690
1700
1710
1720
1730
1740
1750
1760
1770
1780
1790
1800
1810
1820
1830
1840
1850
1860
1870
1880
1890
1900
1910
1920
1930
1940
1950
1960
1970
1980
1990
2000
2010
2020
2030
2040
2050
2060
2070
2080
2090
2100
2110
2120
2130
2140
2150
2160
2170
2180
2190
2200
2210
2220
2230
2240
2250
2260
2270
2280
2290
2300
2310
2320
2330
2340
2350
2360
2370
2380
2390
2400
2410
2420
2430
2440
2450
2460
2470
2480
2490
2500
2510
2520
2530
2540
2550
2560
2570
2580
2590
2600
2610
2620
2630
2640
2650
2660
2670
2680
2690
2700
2710
2720
2730
2740
2750
2760
2770
2780
2790
2800
2810
2820
2830
2840
2850
2860
2870
2880
2890
2900
2910
2920
2930
2940
2950
2960
2970
2980
2990
3000
3010
3020
3030
3040
3050
3060
3070
3080
3090
3100
3110
3120
3130
3140
3150
3160
3170
3180
3190
3200
3210
3220
3230
3240
3250
3260
3270
3280
3290
3300
3310
3320
3330
3340
3350
3360
3370
3380
3390
3400
3410
3420
3430
3440
3450
3460
3470
3480
3490
3500
3510
3520
3530
3540
3550
3560
3570
3580
3590
3600
3610
3620
3630
3640
3650
3660
3670
3680
3690
3700
3710
3720
3730
3740
3750
3760
3770
3780
3790
3800
3810
3820
3830
3840
3850
3860
3870
3880
3890
3900
3910
3920
3930
3940
3950
3960
3970
3980
3990
4000
4010
4020
4030
4040
4050
4060
4070
4080
4090
4100
4110
4120
4130
4140
4150
4160
4170
4180
4190
4200
4210
4220
4230
4240
4250
4260
4270
4280
4290
4300
4310
4320
4330
4340
4350
4360
4370
4380
4390
4400
4410
4420
4430
4440
4450
4460
4470
4480
4490
4500
4510
4520
4530
4540
4550
4560
4570
4580
4590
4600
4610
4620
4630
4640
4650
4660
4670
4680
4690
4700
4710
4720
4730
4740
4750
4760
4770
4780
4790
4800
4810
4820
4830
4840
4850
4860
4870
4880
4890
4900
4910
4920
4930
4940
4950
4960
4970
4980
4990
5000
5010
5020
5030
5040
5050
5060
5070
5080
5090
5100
5110
5120
5130
5140
5150
5160
5170
5180
5190
5200
5210
5220
5230
5240
5250
5260
5270
5280
5290
5300
5310
5320
5330
5340
5350
5360
5370
5380
5390
5400
5410
5420
5430
5440
5450
5460
5470
5480
5490
5500
5510
5520
5530
5540
5550
5560
5570
5580
5590
5600
5610
5620
5630
5640
5650
5660
5670
5680
5690
5700
5710
5720
5730
5740
5750
5760
5770
5780
5790
5800
5810
5820
5830
5840
5850
5860
5870
5880
5890
5900
5910
5920
5930
5940
5950
5960
5970
5980
5990
6000
6010
6020
6030
6040
6050
6060
6070
6080
6090
6100
6110
6120
6130
6140
6150
6160
6170
6180
6190
6200
6210
6220
6230
6240
6250
6260
6270
6280
6290
6300
6310
6320
6330
6340
6350
6360
6370
6380
6390
6400
6410
6420
6430
6440
6450
6460
6470
6480
6490
6500
6510
6520
6530
6540
6550
6560
6570
6580
6590
6600
6610
6620
6630
6640
6650
6660
6670
6680
6690
6700
6710
6720
6730
6740
6750
6760
6770
6780
6790
6800
6810
6820
6830
6840
6850
6860
6870
6880
6890
6900
6910
6920
6930
6940
6950
6960
6970
6980
6990
7000
7010
7020
7030
7040
7050
7060
7070
7080
7090
7100
7110
7120
7130
7140
7150
7160
7170
7180
7190
7200
7210
7220
7230
7240
7250
7260
7270
7280
7290
7300
7310
7320
7330
7340
7350
7360
7370
7380
7390
7400
7410
7420
7430
7440
7450
7460
7470
7480
7490
7500
7510
7520
7530
7540
7550
7560
7570
7580
7590
7600
7610
7620
7630
7640
7650
7660
7670
7680
7690
7700
7710
7720
7730
7740
7750
7760
7770
7780
7790
7800
7810
7820
7830
7840
7850
7860
7870
7880
7890
7900
7910
7920
7930
7940
7950
7960
7970
7980
7990
8000
8010
8020
8030
8040
8050
8060
8070
8080
8090
8100
8110
8120
8130
8140
8150
8160
8170
8180
8190
8200
8210
8220
8230
8240
8250
8260
8270
8280
8290
8300
8310
8320
8330
8340
8350
8360
8370
8380
8390
8400
8410
8420
8430
8440
8450
8460
8470
8480
8490
8500
8510
8520
8530
8540
8550
8560
8570
8580
8590
8600
8610
8620
8630
8640
8650
8660
8670
8680
8690
8700
8710
8720
8730
8740
8750
8760
8770
8780
8790
8800
8810
8820
8830
8840
8850
8860
8870
8880
8890
8900
8910
8920
8930
8940
8950
8960
8970
8980
8990
9000
9010
9020
9030
9040
9050
9060
9070
9080
9090
9100
9110
9120
9130
9140
9150
9160
9170
9180
9190
9200
9210
9220
9230
9240
9250
9260
9270
9280
9290
9300
9310
9320
9330
9340
9350
9360
9370
9380
9390
9400
9410
9420
9430
9440
9450
9460
9470
9480
9490
9500
9510
9520
9530
9540
9550
9560
9570
9580
9590
9600
9610
9620
9630
9640
9650
9660
9670
9680
9690
9700
9710
9720
9730
9740
9750
9760
9770
9780
9790
9800
9810
9820
9830
9840
9850
9860
9870
9880
9890
9900
9910
9920
9930
9940
9950
9960
9970
9980
9990
10000
10010
10020
10030
10040
10050
10060
10070
10080
10090
100100
100200
100300
100400
100500
100600
100700
100800
100900
1001000
1002000
1003000
1004000
1005000
1006000
1007000
1008000
1009000
10010000
10020000
10030000
10040000
10050000
10060000
10070000
10080000
10090000
100100000
100200000
100300000
100400000
100500000
100600000
100700000
100800000
100900000
1001000000
1002000000
1003000000
1004000000
1005000000
1006000000
1007000000
1008000000
1009000000
10010000000
10020000000
10030000000
10040000000
10050000000
10060000000
10070000000
10080000000
10090000000
100100000000
100200000000
100300000000
100400000000
100500000000
100600000000
100700000000
100800000000
100900000000
1001000000000
1002000000000
1003000000000
1004000000000
1005000000000
1006000000000
1007000000000
1008000000000
1009000000000
10010000000000
10020000000000
10030000000000
10040000000000
10050000000000
10060000000000
10070000000000
10080000000000
10090000000000
100100000000000
100200000000000
100300000000000
100400000000000
100500000000000
100600000000000
100700000000000
100800000000000
100900000000000
1001000000000000
1002000000000000
1003000000000000
1004000000000000
1005000000000000
1006000000000000
1007000000000000
1008000000000000
1009000000000000
10010000000000000
10020000000000000
10030000000000000
10040000000000000
10050000000000000
10060000000000000
10070000000000000
10080000000000000
10090000000000000
100100000000000000
100200000000000000
100300000000000000
100400000000000000
100500000000000000
100600000000000000
100700000000000000
100800000000000000
100900000000000000
1001000000000000000
1002000000000000000
1003000000000000000
1004000000000000000
1005000000000000000
1006000000000000000
1007000000000000000
1008000000000000000
1009000000000000000
10010000000000000000
10020000000000000000
100300000

BACKGROUND OF THE INVENTION

1. Field of the Invention

The invention describes a method for upgrading a data stream of

5 multimedia data, which comprises features with textual description.

2. Description of the Related Art

In order to exactly describe e.g. the pronunciation of a text, e.g. for

10 controlling a speech synthesizer, the “World Wide Web Consortium” (W3C) is currently specifying a so-called “Speech Synthesis Markup Language” (SSML, <http://www.w3.org/TR/speech-synthesis>). Within this specification, xml (Extensible Markup Language) elements are defined for describing how the elements of a text are to be pronounced exactly.

15 For the phonetic transcription of text the “International Phonetic Alphabet” (IPA) is used. The use of this phoneme element together with high-level multimedia description schemes enables the content creator to exactly specify the phonetic transcription of the description text. However, if there are multiple occurrences of the same words in different parts of a description text, the 20 phonetic description has to be inserted (and thus stored or transmitted) for each of the occurrences.

SUMMARY OF THE INVENTION

It is an object of the present invention to provide a method of upgrading a
5 multimedia data stream to include text pronunciation information, which avoids
the above-described disadvantage.

It is also an object of the present invention to provide a method, which
enables a more efficient phonetic representation of specific parts or words of
high-level, textual multimedia description schemes.

10 This objective is achieved by means of the present invention in that in
addition to the textual description a set of phonetic translation hints is included.
These phonetic translation hints specify the phonetic transcription of parts or
words of the textual description. The phonetic transcription enables applications
like speech recognition or text to speech systems to cope with special cases
15 where automatic transcription is not applicable or to completely cut out the
process of automatic transcription.

A second aspect of the invention is the efficient binary coding of the
phonetic translation hints values in order to allow low bandwidth transmission or
storage of respective description data containing phonetic translation hints.

20 Known solutions allow the phonetic transcription of specific parts or words
of the description text for high-level multimedia descriptions. However, the
phonetic transcriptions have to be specified for each occurrence of a word or text
part, i.e. if certain words occur more than once in a description text, the phonetic

transcriptions have to be repeated each time. The present invention has the advantage that it permits specification of a phonetic transcription of specific parts or words of any description text within high-level feature multimedia description schemes. In contrast to the state of the art, the present invention permits

5 specification of the phonetic transcription of words, which are valid for the whole description text or parts of it, without requiring that the phonetic transcription is repeated for each occurrence of the word in the description text. In order to achieve this goal, a set of phonetic translation hints is included in the description schemes. These translation hints uniquely define how to pronounce specific

10 words of the description text. The phonetic translation hints are valid for either the whole description text or parts of it, depending on which level of the description scheme they are included. By this, it is possible to specify (and thus transmit or store) the phonetic transcription of a set of words *only once*. This phonetic transcription is then valid for all occurrences of those words in that part

15 of the text where the phonetic translation hints are valid. This makes the parsing of the descriptions easier, since the description text no longer carries all the phonetic transcriptions in-line, but they are treated separately. Further, it facilitates the authoring of the description text, since the text can be generated separately from the transcription hints. Finally, it reduces the amount of data

20 necessary for storing or transmitting the description text.

DETAILED DESCRIPTION OF THE INVENTION

Before discussing the details of the invention some definitions, especially

5 those used in MPEG-7, are presented.

In the context of the MPEG-7 standard that is currently under development, a textual representation of the description structures for the description of audio-visual data content in multimedia environments is used. For this task, the *Extensible Markup Language (XML)* is used, where the Ds and DSs 10 are specified using the so-called *Description Definition Language (DDL)*. In the context of the remainder of this document, the following definitions are used:

• **Data:** Data is audio-visual information that will be described using MPEG-7, regardless of storage, coding, display, transmission, medium or technology.

15 • **Feature:** A feature is a distinctive characteristic of the data, which signifies something to somebody.

• **Descriptor (D):** A descriptor is a representation of a feature. A descriptor defines the syntax and the semantics of the feature representation.

20 • **Descriptor Values (DV):** A descriptor value is an instantiation of a descriptor for a given data set (or subset thereof) that describes the actual data.

• **Description Scheme (DS):** A description scheme specifies the structure and semantics of the relationships between its components, which may be both descriptors (Ds) and description schemes (DSs)

- **Description:** A description consists of a DS (structure) and the set of descriptor values (instantiations) that describe the data.
- **Coded Description:** A coded description is a description that has been encoded to fulfill relevant requirements, such as compression efficiency, error resilience, random access, etc.
- **Description Definition Language (DDL):** The description definition language is a language that allows the creation of new description schemes and, possibly, descriptors. It also allows the extension and modification of existing description schemes.

The lowest level of the description is a descriptor. It defines one or more features of the data. Together with the respective DVs it is used to actually describe a specific piece of data. The next higher level is a description scheme which contains at least two or more components and their relationships.

15 Components can be either descriptors or description schemes. The highest level so far is the description definition language. It is used for two purposes: first, the textual representations of static descriptors and description schemes are written using the DDL. Second, the DDL can also be used to define a dynamic DS using static Ds and DSs.

With respect to the MPEG-7 descriptions, two kinds of data can be distinguished. First, the low level features describe properties of the data like e.g. the dominant color, the shape or the structure of an image or a video sequence. These features are, in general, extracted automatically from the data. On the

other hand, MPEG-7 can also be used to describe high-level features like e.g. the title of a film, the author of a song or even a complete media review with respect to the corresponding data. These features are, in general, not extracted automatically, but edited manually or semi-automatically during production or 5 post-production of the data. Up to now, the high level features are described in textual form only, possibly referring to a specified language or thesaurus. A simple example for the textual description of some high level features is given below.

```
10 <CreationInformation>
    <Creation>
        <Title type="original">
            <TitleText xml: lang="en">Music</TitleText>
        </Title>
    15 <Creator>
        <Role CSName="MPEG_roles_CS" CSTermID="47">
            <Label xml: lang="en">presenter</Label>
        </Role>
        <Individual>
            <Name>Madonna</Name>
        20 </Individual>
        </Creator>
    </Creation>
    <MediaReview>
    25 <Reviewer>
        <FirstName>Alan</FirstName>
        <GivenName>Bangs</GivenName>
        </Reviewer>
```

```
<RatingCriterion>
  <CriterionName>Overall</CriterionName>
  <WorstRating>1</WorstRating>
  <BestRating>10</BestRating>
5 </RatingCriterion>
<RatingValue>10</RatingValue>
<FreeTextReview>
  This is again an excellent piece of music from our well-
  known superstar, without the necessity for more than 180
10 bpm in order to make people feel excited. It comes along
  with harmonic yet clearly defined transitions between pieces
  of rap-like vocals, well known for e.g. from the Kraut-
  Rappers "Die fantastischen 4" and their former chart
  runner-up "MfG", and on the other hand peaceful sounding
  instrumental sections. Therefore this song deserves a clear
15 10+ rating.
</FreeTextReview>
</MediaReview>
</CreationInformation>
```

20 The example uses the XML language for the descriptions. The text in the
brackets ("<...>") is referred to as XML tags, and it specifies the elements of the
description scheme. The text between the tags are the data values of the
description. The example describes the title, the presenter and a short media
25 review of an audio track called "Music" from the well-known American Singer
"Madonna". As can be seen, all the information is given in textual form, possibly
according to a specified language ("de" for German, or "en" for English) or to a
specified thesaurus. The text describing the data can in principle be pronounced

in different ways, depending on the language, the context or the usual customs with respect to the application area. However, the textual description as specified up to now is the same, regardless of the pronunciation.

In order to exactly describe e.g. the pronunciation of the text, e.g. for

5 controlling a speech synthesizer, the “World Wide Web Consortium” (W3C) is currently specifying a so- called “Speech Synthesis Markup Language” (SSML, <http://www.w3.org/TR/speech-synthesis>). Within this specification, xml elements are defined for describing how the elements of a text are to be pronounced exactly. Among others, a phoneme element is defined which allows to specify the 10 phonetic transcription of text parts like described below.

```
<phoneme ph="t&#252; m&#251; to&#28A;"> tomato </phoneme>
```

```
<!-- This is an example of IPA using character entities -->
```

```
15 <phoneme ph="t  m  to"> tomato </phoneme>
```

```
<!-- This example uses the Unicode IPA characters. -->
```

```
<!-- Note: this will not display correctly on most browsers. -->
```

As can be seen, for the phonetic transcription the “International Phonetic

20 Alphabet” (IPA) is used. The use of this phoneme element together with high- level multimedia description schemes enables the content creator to exactly specify the phonetic transcription of the description text. However, if there are multiple occurrences of the same words in different parts of a description text, the

phonetic description has to be inserted (and thus stored or transmitted) for each of the occurrences.

The broad or general concept of the present invention is to define a new DS called “PhoneticTranslationHints” which gives additional information about 5 how a set of words is pronounced. The current Textual Datatype, which does not include this information, is defined with respect to the MPEG-7 Multimedia Description Schemes CD as follows:

```
10      <!-- ##### -->
      <! -- Definition of Textual Datatype -->
      <!-- ##### -->

<complexType name="TextualType">
  <simpleContent>
    <extension base="string">
      <attribute ref="xml: lang" use="optional"/>
    </extension>
  </simpleContent>
</complexType> .
20
```

The Textual Datatype only contains a string for text information and an optional attribute for the language of the text. The additional information about how some or all words in an instance of the Textual Datatype are pronounced is given by an instance of the new defined “PhoneticDecriptionHintsType”. Two 25 solutions for the definition of this new type are given in the following subsections.

The first embodiment of the “PhoneticTranslationHintsType” is given by the following definition:

```
<complexType name="PhoneticTranslationHintsType">
  5   <sequence maxOccurs="unbounded">
    <element name="Word">
      <complexType>
        <simpleContent>
          <extension base="string">
            10 <attribute name="phonetic_translation"
              type="string" use="required"/>
          </extension>
        </simpleContent>
      </complexType>
      15 </element>
    </sequence>
  </complexType>
```

Table I Semantics of “PhoneticTranslationHintsType” Version 1

| Name | Definition |
|--------------------------|---|
| PhoneticTranslationHints | Contains a set of words and their corresponding pronunciations. |
| Word | Single word coded as string. |
| Phonetic_translation | This element contains the additional phonetic information about the corresponding text. For the representation of the phonetic information, the IPA |

| | |
|--|--|
| | (International Phonetic Alphabet) or the SAMPA representation is chosen. |
|--|--|

This newly created type unambiguously gives a connection between words and their appropriate pronunciation. In the following, an example with an 5 instance of the “PhoneticTranslationHintsType” is given which refers to the example discussed before.

<PhoneticTranslationHints>

10 <Word phonetic_translation= "b˜ pÓ miA; n+">

bpm</Word>

15 <Word phonetic_translation= "krŴ r peĢ">

Kraut-Rappers</Word>

<Word phonetic_translation= "em ef g">

MFG</Word>

</PhoneticTranslationHints>

15

With this example of the “PhoneticTranslationHintsType” an application now knows the exact phonetic transcription of some or all words of the text, which is given between the <FreeTextReview> tags in the example discussed before.

20

A second embodiment of the “PhoneticTranslationHintsType” is given by the following definition.

```
5 <complexType name =“PhoneticTranslationHintsType”>
  <sequence maxOccurs=“unbounded”>
    <element name=“Word” type=“string”/>
    <element name=“PhoneticTranslation”/>
  </sequence>
</complexType>
```

10

The semantics of the newly defined “PhoneticTranslationHintsType”, which are the same as in the version 1 described in the previous section, are specified in the following table.

15 **Table II Semantics of “PhoneticTranslationHintsType” Version 2**

| Name | Definition |
|--------------------------|--|
| PhoneticTranslationHints | Contains a set of words and their corresponding pronunciations. |
| Word | Single word coded as string. |
| Phonetic_translation | This element contains the additional phonetic information about the corresponding text. For the representation of the phonetic information, the IPA (International Phonetic Alphabet) or the SAMPA representation is chosen. |

In the following, an example of the “PhoneticTranslationHintsType” Version 2 is given, which refers again to the example discussed before.

```
5   <PhoneticTranslationHints>
    <Word>bpm</Word>
    <phonetic_translation>b&#152; p&#211; mi&#28A; n&#043
    </phonetic_translation>
    <Word>Kraut-Rappers</Word>
10  <phonetic_translation>kr&#372; r&#011; pe&#290;
    </phonetic_translation>
    <Word>MFG</Word>
    <phonetic_translation>em&#001; ef&#005; g&#011;
    </phonetic_translation>
15  </PhonetictranslationHints>
```

With this new definition of the “PhoneticTranslationHintsType” an example of this type consists of the tags **<Word>** and **<PhoneticTranslation>** which always correspond to each other and build one unit that describes a text and its 20 associated phonetic transcription.

The phonemes used in the above-described phonetic translation hints DSs are in general described also as printable characters using UNICODE presentation. However, in general the set of phonemes that is used will be restricted to a limited number. Therefore, for more efficient storage and

transmission a binary fixed length or variable length code representation can be used for the phonemes, which eventually takes into account the statistics of the phonemes.

The additional phonetic transcription information is necessary for a huge number of applications, which include a TTS functionality or speech recognition system. In fact the speech interaction with any kind of multimedia system is based on a single language, normally the native language of the user. Therefore the HMI (the known vocabulary) is adapted to this language. Nevertheless, the words which are used from the user or which should be presented to the user can also include terms of another language. Thus, the TTS system or speech recognition does not know the right pronunciation for these terms. Using the proposed phonetic description solves this problem and makes the HMI much more reliable and natural.

A multimedia system providing content of any kind to the user needs such phonetic information. Any additional text information about the content can include technical terms, names or other words needing special pronunciation information to present it to the user via TTS. The same holds for news, emails or other information, which should be read to the user.

Especially a film or music storage device, which can be a CD, CD-ROM, DVD, MP3, MD or any other device, contains a lot of films and songs with a title, actor name, artist name, genre, etc. The TTS system does not know how to pronounce all these words and the speech recognition cannot recognize such words. If the user, for example, wants to listen to pop music and the multimedia

system should give a list of available pop music via TTS, it would not be able to pronounce the found CD titles, artist names or song names without additional phonetic information.

If the multimedia system should present (via text-to-speech interfaces 5 (TTS)) a list of the available film or music genres, it also needs this phonetic transcription information. The same also holds for the speech recognition to better identify corresponding elements of the textual description.

Another application is the radio (via FM, DAB, DVB, RDM, etc.). If the user wants to listen to the radio and the system should present a list of the available 10 programs, it would not be possible to pronounce the programs, because the radio programs have names like “BBC”, or “WDR”. Others have a name using normal words like “Antenne Bayern” and some names are a mixture of both, e.g. “N-Joy”.

The telephone application often provides a telephone book. Even in this 15 case without phonetic transcription information the system cannot recognize or present the names via TTS, because it does not know how to pronounce it.

So any functionality or application which presents information to the user via TTS or which uses a speech recognition needs a phonetic transcription for some words.

20 Optionally it is possible to transmit the reference on any given alphabet, which is used to represent the phonetic element.

The translation hints together with the corresponding elements of the textual description can be implemented in text-to-speech interfaces, speech

recognition devices, navigation systems, audio broadcast equipment, telephone applications, etc., which use textual description in combination with phonetic transcription information for search or filtering of information.

The disclosure in German Patent Application 01 100 500.6 of January 9, 5 2001 is incorporated here by reference. This German Patent Application describes the invention described hereinabove and claimed in the claims appended hereinbelow and provides the basis for a claim of priority for the instant invention under 35 U.S.C. 119.

While the invention has been illustrated and described as embodied in a 10 method of upgrading a data stream of multimedia data, it is not intended to be limited to the details shown, since various modifications and changes may be made without departing in any way from the spirit of the present invention.

Without further analysis, the foregoing will so fully reveal the gist of the present invention that others can, by applying current knowledge, readily adapt it 15 for various applications without omitting features that, from the standpoint of prior art, fairly constitute essential characteristics of the generic or specific aspects of this invention.

What is claimed is new and is set forth in the following appended claims.